63-3-4

AMRL-TDR-63-8

INFORMATION STORAGE AND RETRIEVAL: A SURVEY

George W. Barnard, Captain, USAF, MC
Carl Abbott

TECHNICAL DOCUMENTARY REPORT NO. AMRL-TDR-63-8

JANUARY 1963

BIOMEDICAL LABORATORY
6570th AEROSPACE MEDICAL RESEARCH LABORATORIES
AEROSPACE MEDICAL DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO

403 889

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Qualified requesters may obtain copies from ASTIA. Orders will be expedited if placed through the librarian or other person designated to request documents from ASTIA.

Do not return this copy. Retain or destroy.

Stock quantities available at Office of Technical Services, Department of Commerce, $0.75.

Change of Address

Organizations receiving reports via the 6570th Aerospace Medical Research Laboratories automatic mailing lists should submit the addressograph plate stamp on the report envelope or refer to the code number when corresponding about change of address.

27-663, 700, 5-8-63

FOREWORD

This study was performed by the Psychophysiological Stress Section, Biophysics Branch, Biomedical Laboratory. This study began in June 1962 and was completed in September 1962. Carl Abbott was associated with the Psychophysiological Stress Section during this period.

ABSTRACT

This report surveys the field of information retrieval, the methods of indexing and storing the vast number of scientific documents which have been produced in recent years. Information retrieval utilizes coordinate indexing—that is, listing documents under all the topics they contain and searching for them by two or more terms. There are two principle types of indexing: one using a predetermined list of terms into which all documents must be fitted and the other allowing free choice of the terms found in the documents themselves. Elaboration of these methods and the difficulty of developing a list of indexing terms are also discussed. An information retrieval system may consist of an index only, an index with an abstract, or an entire document with an index. The mechanical equipment used may range from punched cards through IBM cards to complex computers and microphotographic systems. The experiences of various organizations with different combinations of equipment and methods are discussed.

PUBLICATION REVIEW

This technical documentary report has been reviewed and is approved.

Jos M Quashnock

JOS. M. QUASHNOCK
Colonel, USAF, MC
Chief, Biomedical Laboratory

INFORMATION STORAGE AND RETRIEVAL: A SURVEY

INTRODUCTION

At the present time, human knowledge is increasing at a fantastic rate. More scientific research has probably been done in the past 20 years than in the preceding 200 years. Between 50,000 and 100,000 technical journals are now being published in over 60 languages. Languages such as Russian, Chinese, and Japanese have far surpassed German, French, and Italian in number of articles. No one knows exactly how many articles are published, but the number lies between 1 million and 5 million per year. In some fields, the amount of research doubles within 10 years. The problem is simple: how to extract one particular piece of information from this mass of material. As Vannevar Bush said, "The summation of human experience is being expanded at a prodigious rate; the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships."

History

The effort to bring this means up to an atomic-age standard is the science of information retrieval. As a recognizable, distinct science, it is no more than 15 years old, but it now absorbs about one-eighth of the nation's research and development budget. Although libraries and classification systems since the time of Aristotle have performed the function of identifying the location of a specific bit of information, conventional systems are inadequate under the massive burden of recent years. Since World War II, a great amount of study has been applied to the development of the most logical and foolproof methods of indexing materials. With the development of electronic computers, much attention has been given to the mechanics of storage and retrieval. Almost all the large electronics corporations are active in the field, and many companies devoted exclusively to the science have sprung up. Furthermore, many large industries and government agencies (including the Patent Office, ASTIA,* and the CIA†) have installed systems of varying complexity.

---

*Armed Services Technical Information Agency (ASTIA), Arlington Hall Station, Arlington 12, Virginia

†Central Intelligence Agency (CIA), 2430 E. Street, N.W., Washington, D.C.

Scope of Problem

Information retrieval is a many-faceted subject. Basically, it seems very simple, a method of indexing plus a method of making use of this index. However, the ramifications seem endless. The method of "making use" leads into the field of computers and electronic processing with its elaborate technology, abstruse mathematics, and applications of physics. On the other hand, indexing leads directly into logic and from there into linguistics and semantics, systems of mathematics, and ultimately into investigation of the processes of human thought.

In this report we limit our discussion to those aspects of information retrieval most directly of interest to the user: types of indexing and methods of mechanization. We make no attempt either to probe the logical-linguistic-mathematical-philosophical depths that underlie indexing or to make this, in any way, a discussion of electronic data processing. Neither do we try to cover everything in a complete, technical manner. This report is divided into three major sections: method of indexing or the logical aspect, method of storage and retrieval or the mechanical aspect, and a discussion of some working systems.

METHOD OF INDEXING

Fundamentally, an index is a means of identifying an item and assigning it a fixed quality so that it can be readily found. Thus, if all books on a certain subject were given a particular color of binding, one would have a type of indexing system. In a card catalog, three qualities—author, title, and subject—are used to place a book. The Dewey decimal system is a method of indexing by hierarchical classification; that is, a broad subject is broken down into smaller and smaller subdivisions until there is a unique grouping of divisions for each book. The more items there are to be indexed, the more complex the system must be.

Coordinate Indexing

Almost all work in information retrieval involves a particular type of indexing: coordinate indexing. Coordinate indexing is a means used for locating an item by the intersection of two or more terms. Each item is identified by several terms which apply to it—for example, "bridge supports, aluminum, special stresses, suspension bridges." With a card catalog, a searcher would have to look through all the articles under "aluminum" plus all those under "bridge supports" to make sure he found this item, and might still miss it if it were listed under "structural metals." In a hierarchical classification, he would have to trace through "architecture → bridges → bridge supports → special stresses on aluminum supports," "metals → aluminum → special stresses on aluminum → aluminum bridge supports," and all the other possible series.

However, with a coordinate system, only one search need be made. The system can best be visualized as a rectangular graph:

Concrete				
Aluminum		X		
Steel				
	Roads	Bridges	Houses	Skyscrapers

Figure 1. Illustration of Coordinate Indexing

Only those items which are listed under both "aluminum" and "bridges" will be retrieved. If this is not specific enough, additional terms can be used. Those items which lie at the intersection of all the items (that is, are listed under each term) will be retrieved. Thus, supposing an adequate mechanical system to do the searching, the only requirements for retrieval of all relevant items are the listing of an item under all applicable terms and the use by the searcher of the terms most accurately describing his request.

Classes of Systems:

All coordinate systems can be described as belonging to one of two classes, term-on-item and item-on-term. (An item is any article, book, document, etc., to be indexed and a term is any word used to describe an item.) Though the actual mechanics may be much more complex, they are best explained by thinking of edge-notched cards. In a term-on-item type, there is one card for each document, containing a punchhole for each term under which items may be listed. In the appropriate places, all the terms describing that item are punched, or the terms are indicated on the item card. When all the cards are searched, those punched for certain terms fall out. In an item-on-term type, there is a card for every term, and each card has a punch space for every document in the file. In search, only the cards for the relevant terms need be used. When superimposed, it is easy to see those punches which, appearing on all of them, indicate relevant documents.

Development of Vocabulary

The greatest problem of coordinate indexing is the development of a suitable code or list of terms. A system is obviously worthless if documents are listed under irrelevant terms, or not under all relevant ones, or if a user runs a search using terms different from those the items are listed under. There has to be some standard for the choice of terms by indexer and user.

Unlimited Terms:

It is possible to have a system which has no limitation or control of the terms used, but instead the most relevant ones are freely picked from the document itself, without regard to whether they have been used before. The advantage of this method is that the terms that most accurately describe an item are used, being the words of the document itself. Among the disadvantages are that such a system must be item-on-term, to allow for the unlimited number of terms, and that the user has no idea whether he is using the correct terms to retrieve an item, removing the accuracy and preciseness that make coordinate indexing valuable.

Confined Choice of Terms:

Most kinds of systems use some sort of predetermined vocabulary, or lists of terms, which may be used to index documents. The list of terms may be of any size, depending on the number of items and the size of the organization using it. Not only does a limited vocabulary have the advantage of being applicable to both kinds of systems and of assuring the user that there are no terms which he has overlooked, but it also has the advantage of being adaptable to any sort of specialized service and of being easily controlled and modified. With a definite number of terms, those which are inapplicable to a specific field can be eliminated and special ones developed. Seldom used terms can be discovered and replaced by more important ones. Any unsanctioned terms can be discovered and eliminated. These advantages far outweigh the disadvantage that some items may be forced into headings not exactly suited for them.

Development of Term List:

Many limited systems use authority lists, or thesauri. These are, in effect, dictionaries which list terms which are acceptable synonyms for unacceptable words. These books enable the indexer to place a doubtful document under the right headings and a user to convert his question into useful terms.

While an unlimited list of terms is merely taken from the documents, a definite list must be developed before any use may be made of it. Sometimes this list can be developed from an already existing system of hierarchical classification or subject headings. (The latter was done by ASTIA, which reduced 70,000 subject headings to 9,000 terms.) Another method is to pick a list of terms based on a knowledge of the field to be covered, modifying it after it has been in use. A third method is to take a sample group of documents and index them freely. All the terms used are then gone over and synonyms eliminated, seldom used terms combined or eliminated, and overused ones broken down. Whatever system is used, it will certainly require extensive modification on the basis of experience.

Methods of Coordinate Indexing

There are many methods which can be used to achieve a coordinate index. Most of them have been put into active practice, and none of them can be judged except on the basis of actual performance in the established system. However, they are not specialized or adapted for only one organization, but are general theories applicable to any problem. The following is a list and explanation of some of the more important methods.

Uniterms:

Uniterms, a system developed by Documentation, Inc., a subsidiary of Benson-Lehner Corporation, * is perhaps the simplest method. "Uniterms" are the significant words, or keywords, contained in a document. Each document must be listed under all the keywords it contains. Thus, it is a system of free indexing, as discussed in the preceding section. Its main advantage is depth and accuracy; that is, the insurance that important parts of an item have not been left unindexed because they deal with a rarely encountered topic.

Descriptors:

Descriptors, a term originated by Calvin Mooers of Zator Company, † is another name for limited terms, a vocabulary of general, inclusive terms for a particular field. Although very easy to use, a system of this type cannot hope to have as great a depth or accuracy as uniterms. Any need for greater preciseness or an increase in the subject area requires a parallel increase in the number of descriptors. However, almost all requests can be handled with such a system. The number of items which would not be retrieved when listed under 10 or 12 descriptors, but would be found by complete free listing, is very small.

One additional problem with a descriptor method is that every organization must develop its own list of terms to fit its personal needs. While the thesauri of other organizations (either those in the same field or ones with broad coverage, such as ASTIA) may be studied, individual work and development is essential. Although a satisfactory development of vocabulary results in an index well fitted for the organization, a poorly planned development might render the entire system unsatisfactory.

*Benson-Lehner, 14761 Califa, Van Nuys, California
†Zator Company, 140½ Mount Auburn, Boston, Massachusetts

Roles and Links:

Roles and links are methods of increasing the preciseness of the index. In effect, they are a type of grammar, showing relationships between terms, perhaps [(aluminum, stress) - (bridge)], where aluminum and stress are very closely related, and both terms connected somewhat less closely to bridge. Roles show the relationships of words in context. For example, the use of a role indicator with the words "heat" and "water" would show whether "water for heating" or "heating of water" was meant. The implementation of such devices is very complicated, for they require an electronic computer for proper use. However, in a medium-sized system, they provide the most sophisticated and exact indexing available.

Permuted Titles:

Permuted titles are a quick, simple index developed by the Chemical Abstracts Service.* The title of a document should give a rough idea of the contents. Therefore, the title is listed alphabetically under each important word it contains. For example, the following title could be found under these three listings:

Corrosion of Magnesium by Sea Water

Corrosion of Magnesium by Sea Water

Corrosion of Magnesium by Sea Water

This is obviously not a method with pinpoint accuracy, for a title can mention only the broadest topics (and even then may make a poor choice), but it can be produced quickly and inexpensively without needing to process the entire article.

Special Codes:

In some fields, such as chemistry, special code systems have been devised to index certain types of information. For instance, a chemical compound could be coded as ABCDE, with each letter representing some chemical structure. This particular compound would be retrieved if one were to require all reports containing compounds having, for example, structures A, B, D, and E together. This, of course, is not a complete method of identifying reports, but a system of sub-indexing some of the main indexing terms.

Text Searching:

Computer text searching has been developed by the General Electric (GE) Company† and used at Western Reserve University.‡ It is a special variety of term-on-item indexing. From each item, an abstract of 15 or 20 words is taken. When a request is received, the machine scans each abstract and selects those having the right combination of terms. This procedure differs from a normal term-on-item search in that the machine, instead of scanning indexing terms, actually reads a summary of the document and decides whether it is relevant. It may also be possible to scan entire articles in this manner, although the amount of time needed might be prohibitive. The principle disadvantage of this method, aside from the difficulty of producing the abstract, is the basic difficulty with any such use of computers: the statistically based program cannot have completely adequate sensitivity.

---

*The Chemical Abstracts Services, The Ohio State University, Columbus 10, Ohio
†General Electric Company, Computer Division, 13430 N. Black Canyon Highway, Phoenix, Arizona
‡Western Reserve University, Center for Documentation and Communication, Cleveland, Ohio

Computer Abstracting and Indexing:

Computer abstracting and indexing is not a method, but a means, of indexing. Most attempts at computer indexing have involved picking words which appear most often, or which appear in conjunction with certain other cue words (such as necessary, large, likely) often associated with significant words. This is free indexing based on statistical counts. Research so far, mostly done by IBM* and Ramo-Wooldridge,† has failed to achieve satisfactory results. Abstracting would probably be a compilation, by some method, of the words chosen for indexing.

METHOD OF STORAGE AND RETRIEVAL

The method of storing and retrieving materials, or the mechanical aspect of information retrieval, is equal in importance to the method of indexing. The two aspects are mutually interdependent, and it is as pointless to have an elaborate computer system without a satisfactory index as to have an indexing system without means of making use of it. In recent years, the means of "making use" have grown tremendously with the development of electronic computer systems. The index need no longer be limited by the ability of men to manipulate it, but only by the cost of the system and the depth required.

Quantity of Storage

The first problem in this area is the amount of material to be stored. In most hand-manipulated indexes, such as a card catalog, only the name of the item is listed, making a pure index, or means of identifying the item. In effect, it produces a bibliography of relevant material. The searcher must then retrieve the actual document from its place of storage. Though the storage is usually in conjunction with the index, this procedure requires two steps and a consequent loss of time. In many systems, both the index and the storage may be mechanized (perhaps a computer and a microfilm file). In this case, it is possible that more time may be used in the movement from one section to another than in the actual searching of index and files. On the other hand, a two-stage system may be much less complex and expensive than a one-stage, since an index alone can be made to occupy a very small space. Where the library of items is small enough to make a separate search feasible, a simple index may be the most economical method.

If a one-step system is preferred, either abstracts or entire documents may be stored with the index. The complexity and expense of the mechanical systems increase with the amount of material stored. Where an index alone may be manipulated by hand, placed on IBM cards, or placed in a small section of a computer, storage of a large number of whole documents would necessitate use of a specially designed system.

Storage of abstracts is perhaps more useful. If the system is small, they can be typed directly onto edge-notched cards; a small magnetic computer would be able to handle larger collections. When a search is run, the searcher receives not only the name of the item but also a brief abstract. If the abstract contains information of interest to him, he can then obtain the actual document and study it at leisure. This method assures that the searcher knows exactly what is available to him. On the other hand, he is not forced to go through entire documents to determine their relevancy. Systems are now being developed where the abstract and the index are one and the same thing. The machine searches a specially prepared abstract rather than a special index. If relevant, this abstract itself may be displayed to the sender.

---

*IBM Data Processing Division, 112 E. Post Road, White Plains, New York
†Thompson Ramo Wooldridge, Inc., 8433 Fallbrook Ave., Canoga Park, California

Even more complex systems have the entire item stored along with the index. When the search is run, all the indicated documents are retrieved and presented to the user. This method requires quite elaborate machinery, the documents being stored on magnetic tape or microfilm. While this is the fastest method of retrieval, and greatly reduces the storage space, it is much too expensive for all but large collections where speed is a main requirement. To overcome the possibility of retrieval of a mass of redundant or inapplicable items, the useful system would contain both abstracts and documents. The abstracts would be displayed to the searcher, who would then select the most useful documents.

Complexity of System

The second major problem of mechanization, parallel to that of the amount of material to be stored, is the degree of elaborateness or complexity of the machinery used. There are a great many possibilities in all the phases of mechanization. For input, one can choose from several types of punch cards, microfilm or other camera devices, and magnetic tape. In some systems, such as cards and microfilm, the input device is also the storage device. In others, such as magnetic tape, the input and storage utilize different means. The output and reproduction can also take many forms: (a) the original punched card may be returned; (b) there can be a printout, as from a tape computer; (c) there can be retrieval of filmed information for special viewing or large-size reproductions of greatly reduced film images. In general, a tremendous number of ways to store information has been developed, and each method has certain advantages.

Manually Operated:

The simplest type of system is, of course, manually operated. That is, one in which the machines used are extensions of the human being who uses them, rather than entities in their own right. This type of system commonly makes use of some sort of punch cards. The best known are McBee* Keysort cards, paper cards with coded holes along the edges. When the cards are punched, notches are left in the edges, and when a metal rod is inserted in a particular hole, those cards notched in that position will fall out. Zator Corporation has also developed this type of card. Another type uses holes through the body of the card. The IBM card, with which most people are familiar, is of this type. Jonker Business Machines, Inc.,† has developed the Termatrex card. There is a plastic card for each term in the index and a code space in the card for each document in the file. If a document is to be listed under a certain term, a hole is drilled through the term card in the space for that document. In retrieval, only the cards for the desired terms are used. They are superimposed and a light is shined at them. The light will show through only in the spaces where an item has been indicated on each card, giving a quick bibliography of materials containing all those terms.

Cards such as IBM and Termatrex, which are filled with code spaces, have the advantage of allowing fairly deep indexing, since they have a great many possible code combinations. They have the disadvantages of requiring equipment such as special punch and sorting machines and of providing only an index. Conversely, McBee cards allow space for only a limited index, but can be used very cheaply and provide space for an abstract.

Computers:

The next step in complexity is the use of simple computers. This use can involve a small computer used for information retrieval alone or a larger one with a small section set aside for information retrieval. Using an electronic computer is a tremendous leap over a manually operated system. Almost any index code for almost any volume of material is possible.

---

*Royal McBee, 850 Third Ave., New York 22, New York
†Jonker Business Machines, Inc., 26 North Summit Ave., Gaithersburg, Maryland

The computer can be adapted to an index only, with the material stored elsewhere. In this case, any type of index can be used. Where a manually operated system will be limited in the number of terms by the available space, a computer offers practically unlimited space. However, this may not necessarily be an advantage, because any index may become so complex as to be unmanageable, and it may become so elaborate that it serves no useful purpose. With a computer, there might be temptation to use more elaborate indexes than would be practical.

Adaptation of Computers. —Even more useful, however, is the adaptation of computers not only to indexing but also to storage. The advantages of speed of retrieval, great volume of storage, and elimination of an extra step are obvious, as are the problems of expense and of reproduction and output. Any magnetic tape computer can be adapted to this use. For example, the GE 225 system uses punch cards for input of information, magnetic tape for storage, and punched tape or printout for output. A system of this sort, of course, differs very little from a normal use of electronic data-processing machines, since it utilizes equipment designed for general use.

Specially Constructed Computers. —A step higher in elaborateness are systems designed with the needs of information retrieval in mind, especially those of adequate indexing and reproduction of material. In the field of computers, many companies have produced machines with special characteristics. For example, RCA* has developed a Video File System, a system in which magnetic tape stores images rather than symbols. These images can then be converted directly into reproductions of the stored material. Recordak's† Dacom is a similar system. Magnetic tape used to store the material is scanned by the machine and a microfilm record is produced from the coded information on the tape. Many other systems of equal complexity have been developed to try to solve the problem of modifying a normal computer to do more than merely regurgitate code symbols.

Photographic:

The most involved systems are those built around storage of reduced-image records. Some, such as Avco's‡ Verax, have separate storage and indexing units. This procedure loses the value of a one-step operation, but has the advantage of very high-quality reproduction. NCR's** Photochromic system is another two-step means using a special image-reducing technique which does not inhibit later production of high-quality reprints. Perhaps the climax of the developed systems is Eastman-Kodak's†† Minicard. Documents are photographed onto small chips of microfilm. The document occupies about two-thirds of the space on the film, while the other third is used for an indexing code. When searched, the system yields the pieces of microfilm with the required code, which are then ready for viewing or copying. This system is the purest form of storage-retrieval, with the indexing and storage not only done in the same machine but on the same piece of material. It is, in effect, a refinement of the method of a McBee card with a typed abstract.

---

*Radio Corporation of America (RCA), 30 Rockefeller Plaza, New York 20, New York
†Recordak Corporation, 415 Madison Avenue, New York, New York
‡Avco Corporation (Crosley Division), 1329 Arlington, Cincinnati, Ohio
**National Cash Register Company, Patterson and Steward St., Dayton, Ohio
††Eastman Kodak Company, 343 State Street, Rochester 4, New York

WORKING SYSTEMS

In this section we discuss a few of the information retrieval systems now in use. We try to show how some of the special techniques, both of indexing and of mechanization, have been applied in working systems and how successful these systems have been.

General Characteristics

Most of the working systems on which information is available have several common characteristics.

First, they are large systems, indexing large collections for large organizations. There are several probable reasons for this. Generally, a large-scale system, being a more complex job of engineering, is more interesting to those in the field. A large organization has more money to pay for an expensive installation. A large organization has better means of publicity about its accomplishment. In spite of these factors, more attention should be given to the small systems.

Second, most of the systems are based on electronic computers. Important reasons for this are that a large system requires a computer for quick and easy manipulation, and, at the time of installation, many of the large organizations were already using a computer which could be easily adapted for information retrieval.

The third characteristic is of a different type than the other two. Most of the systems use a special code or indexing method developed to fit the particular needs of the organization. To attempt to use a general, previously developed code is unsatisfactory. As the code would not fit the needs of the organization, the value of the system would be reduced.

Varieties of Coding

The following are examples of different coding methods.

Personal Term List:

Smith, Kline, and French Laboratories* developed a personal term list to take care of a specific problem, 'a great number of reports concerned with one drug, chlorpromazine. Since it was hoped that the system developed would be applicable not only to chlorpromazine, but to any drug, an index was sought that would be specific enough for the one drug, but general enough to be used on any other. The list of index terms which was decided upon has two divisions. There are about 150 terms which list all possible sites of action of the drug—organs, tissues, hormonal activities, etc. There is also a group of descriptive words which can be applied to any term in the first list. These include the experimental subject, the type of study, special circumstances, and type of reaction. Not only is this indexing code easy to use and conducive to a high uniformity of indexing, but it can be expanded at any time, if necessary.

---

*Smith, Kline and French Laboratories, 1530 Spring Garden, Philadelphia, Pennsylvania

A much smaller vocabulary has been developed by the National Conference on Social Welfare,* one of the few small organizations reported on. It was designed to index a fairly small number of frequently used reports and uses hand-punched cards. It consists of 2 sections, 50 descriptive terms (such as Minority Groups and Rehabilitation), and 12 categories (such as Methods, Age Groups, and Purpose). Because of the limited number of terms, the domain of each was defined as exactly as possible. Within its limits (those of size and exactness), the system has performed satisfactorily.

Unlimited Vocabulary:

The GE Flight Propulsion Center† indexes 60,000 reports under approximately 7,000 terms. Although both mechanically and theoretically possible, the list of terms is unlikely to grow much larger. Since the reports indexed are all in the same general area, the most useful terms have already been used. The burden of the system lies on the searchers, who must formulate the problem in the terms most likely to retrieve the necessary information.

The DuPont Corporation‡ also uses an unlimited index on a much smaller scale. Two research departments have indexed their reports to a great depth. Each report is described by approximately 65 terms. The entire index is printed and bound in one volume. The system apparently works very well, for DuPont has tried to keep it a guarded secret. However, it covers only a small number of reports (2,000) in one field and is used by persons very familiar with the possibilities of the field. Also, it was developed with no regard for cost.

Both of these unlimited systems cover a specific field and rely on a searcher who can use a knowledge of the subject area to formulate the question in the correct search terms.

Permuted Titles:

There have been many applications of a permuted title index in recent years. Chemical Abstracts Service publishes Chemical Titles, a semimonthly listing of all the new reports in the field. Coming out several months before the abstracts themselves, the listing enables a researcher to discover the reports covering general areas of interest to him. Bell Telephone Laboratories** has also implemented a permuted title index to internal reports, and feels that the major disadvantages of such an index are the possible inaccuracy of titles and the scatter of subject headings throughout the alphabet. Advantages are the speed of production, portability, and the general clerical nature of production, not requiring a highly trained indexer.

Limited Vocabulary with Thesaurus:

Probably the largest reported information retrieval system is that of ASTIA. In both size of collection—300,000 items—and range of subject matter, it dwarfs most other systems now in use. Because of the size, using anything but a limited and controlled vocabulary would be impossible. Although ASTIA uses a personal vocabulary list, it differs from smaller systems in that the list must be able to take care of any area. For this reason, the system is not so fine and

---

*National Conference on Social Welfare, 22 W. Gay St., Columbus 15, Ohio
†General Electric Flight Propulsion Center, Mill Creek Expressway, Evendale, Ohio
‡E. I. DuPont, de Nemours and Company, Wilmington 98, Delaware
**Bell Telephone Laboratories, 463 West Street, New York 14, New York

precise as for smaller collections, and lacks intercrossing classes of terms. Instead, 19 broad areas (such as Chemistry and Aeronautics) are broken down into 9,000 smaller terms used to list the documents. The terms are listed in the Thesaurus of Descriptors, which acts as a dictionary, giving the scope of each term (as, "Digestive System, includes: Alimentary Canal, Gastro-Intestinal Tract") and the acceptable synonyms (as, "Catapult Seats use Ejector Seats"). The system, especially the Thesaurus, has been under much criticism for being both too cramped and not adequately precise. However, it seems to satisfy the needs of ASTIA quite well, and is not so unwieldy as it might seem.

Roles and Links:

While using roles and links, a means of showing relationships between terms, is the most refined of indexing methods, they can take a very simple form. The ESSO Research and Development Company* uses a rudimentary system of roles and links in its index to internal research reports. In addition to two groups of descriptors, major and minor, there are three coded role indicators applied where needed: Reagent, Intermediate, Final Product. There is also linking of certain related terms by special symbols. For example, where compound I is being searched for with terms A, B, and C, and compound II with X, Y, and Z, the two groups of terms could be tied together so that the search does not yield compounds A, B, Z, or X, Y, C.

A more elaborate system is Western Reserve University's indexing of literature for the American Society of Metals.† Using the same theory as above, a much more complex code system has been evolved. Though much too complex to explain here, its components are the basic index terms (mainly names of materials, properties they possess, or processes they undergo), indicators which tell what role the terms play in the context (for instance, acted upon), and links or "punctuation symbols" used to group the terms with proper relationships.

A role and link coding has the advantage of great accuracy and preciseness. However, such a method is by no means necessary to a good system, and might be inapplicable to some problems.

Levels of Mechanization

This section covers some of the organizations which have already been discussed. However, the focus is on the type of machinery used rather than on the index itself.

Hand Search:

The National Conference on Social Welfare uses a hand-searched, edge-punched card system manufactured by the Zator Company. It is very similar to the McBee Keysort Card System in methods of coding identifying information, such as author's name and 65 holes for code terms. This system was used primarily because of the size of the organization. Where only a small number of documents is to be indexed, and special machinery may be prohibitively expensive, the disadvantages of limited number of code terms and slow search rate are not readily apparent. The chief problem of the system has been poor choice of code terms, rather than the mechanics.

---

*Esso Research and Engineering Company, Technical Information Division, Linden, New Jersey
†American Society for Metals, Route 87 Metals Park, Novelty, Ohio

IBM Cards:

Many organizations use IBM punch cards and card sorters. This method, used by ESSO Research and Development Company, Smith, Kline, and French Laboratories, and others, has the disadvantages of lacking the unlimited capacity and speed of a computer and yet requiring some expensive equipment. However, when compared to a hand search, it has a great capacity, for an IBM card can be coded with a great amount of information. Assigning code values to combinations rather than single punches, Smith, Kline, and French has obtained several thousand terms. For any moderate-sized system, this has proved quite adequate.

Computers:

The amount of information in the largest collections and the complexity of coding in the most involved indexes fall far short of the capacity of some computer systems. Western Reserve University uses a GE 225, the GE Flight Center an IBM 7090, and ASTIA a Univac solid state, and almost as many other computers have been successfully used as have been produced. Though there might be difficulty on some machines in storage of a great mass of information or in time of search with a too elaborate index, there exists no system too large for some computer to handle.

Microfilm:

Unfortunately, there is very little information available on the success of microfilm systems. However, a Minicard system has been installed in the Pentagon and is apparently working satisfactorily.

CONCLUSIONS

An almost unlimited number of combinations of indexes and machinery is now available. The problem is less that of finding a workable system than that of finding the best system for a particular need.

In indexing, the choice does not lie between different types of indexes, such as hierarchical versus coordinate, for it is generally agreed that a coordinate type is far superior. Instead, the choice lies with the amount of freedom to be allowed in the selection of terms. In a small system, a limited number of terms might be satisfactory; in a larger one, it might be overly restrictive. Conversely, while an unlimited choice might be useful for a large collection, it could very well be superfluous effort on a small one. Furthermore, the choice depends not only on the size of the system, but also on the type of use and the type of personnel available to do the indexing and searching. There is no point in having an index that is deeper than necessary. The best indexing system can be made useless if its human operators are not able to handle it properly. There is, therefore, no best system, and no worst, but only one better suited to a particular situation.

The mechanical system is chosen primarily by two factors—expense and adaptability to the index type. The cheapest machinery is, of course, punch cards, then IBM cards and sorters, then adapted computers, then specially designed systems. Each successive type has some advantage over its predecessor in volume of storage, depth of index, or speed of search. The capacity of electronic-based systems, while still finite, is beyond the limit of human need.

However, one should be very careful not to choose a system too complex for his needs. Though the temptation is great to have the best available, money spent on a computer for an organization whose needs could be served by McBee cards or IBM cards would be better invested in the training of personnel to use the system. Since the purpose of information retrieval is to serve people, it can only be successful if it is used often and correctly.

The primary conceptual demand of a data storage-retrieval system is that it be a dynamic functional enterprise. With a good index system, proper coordination, and knowledgeable researchers doing the abstracts, the rewards are expected to be enormous for a vast number of people.

BIBLIOGRAPHY

1. Armed Services Technical Information Agency, Armed Services Technical Information Agency Corporate Author List, Arlington, Virginia, April 1962.

2. Armed Services Technical Information Agency, Automation of ASTIA - 1960, Arlington, Virginia, AD 247000, December 1960.

3. Armed Services Technical Information Agency, Documentation, a Report Bibliography, Arlington, Virginia, AD 267000, December 1961.

4. Armed Services Technical Information Agency, Guidelines for Cataloging and Abstracting, Arlington, Virginia, August 1959.

5. Armed Services Technical Information Agency, Guidelines for Using ASTIA Descriptors, Arlington, Virginia, February 1961.

6. Borden, W. A., W. Hammond and J. H. Heald, Automation of ASTIA - 1959, Armed Services Technical Information Agency, Arlington, Virginia, AD 227000, December 1959.

7. Bello, F. "How to Cope with Information," Fortune, September 1960.

8. Bawker, K., E. J. Lucas, L. H. Martain, C. Phaneuf, Technical Investigation of Elements of a Mechanized Library System, Avco Corporation, Cincinnati, Ohio, January 1960.

9. Day, M. S., and I. Lebow, "New Indexing Pattern for Nuclear Science Abstracts," American Documentation, Vol XI, 1960.

10. Dyson, M., "Closing the Gap in Chemical Documentation," Chemical and Engineering News, Vol 38, May 9, 1960.

11. Dyson, M., "Current Research at Chemical Abstracts," Jour. of Chemical Documentation, Vol 1, p 24, 1961.

12. Dyson, M., and E. Riley, "Mechanized Storage and Retrieval of Organic Chemical Data," Chemical and Engineering News, Vol 39, November 20, 1961.

13. Hayne, R., and F. Turin, "Machine Retrieval of Pharmacologic Data," Advances in Documentation and Library Science, Vol II, Interscience Publishers, Inc., New York, 1957.

14. Heald, J. H., H. Rehbock, D. A. Beobs, M. Brooks, P. M. Klinefelter, P. H. Klingbiel, J. S. Moats, J. V. Philbrick, Thesaurus of ASTIA Descriptors, Armed Services Technical Information Agency, Arlington, Virginia, May 1960.

15. Jahoda, G., "The Development of a Combination Manual and Machine-Based Index to Research and Development Reports," Special Libraries, February 1962.

16. Jahoda, G., M. D. Schoengold, and T. J. Devlin, "A Machine-Based Index to Internal Research and Engineering Reports," Jour. of Chemical Documentation, Vol 1, p 91, 1961.

17. Jaster, J., B. Murray, and M. Taube, The State of the Art of Coordinate Indexing, Documentation, Inc., Washington, D. C., February 1962.

18. Kent, A., "Documentation," Library Trends, Vol IV, No. 27, October 1961.

19. Kent, A., "Documentation and Communication Research," Wilson Library Bulletin, June 1961.

20. Kent, A., "Mechanized Searching and Correlation of Scientific Knowledge," General Semantics Bulletin, Nos. 26-27, 1960.

21. Kent, A., "Resolution of the Literature Crisis in the Decade, 1961-1970," Research Management, Vol V, No. 1, 1962.

22. Klingbiel, P.H., Language Oriented Retrieval Systems, Armed Services Technical Information Agency, Arlington, Virginia, AD 271600, February 1962.

23. Longnecker, H.C., "The Role of a Science Information Department in Index Research and Development," Jour. of Chemical Education, Vol 33, December 1956.

24. Longnecker, H.C., "Staffing an Information Group," presented at Symposium on the Administration of Technical Information Groups, American Chemical Society, Chicago, September 9, 1958.

25. National Science Foundation, Current Research and Development in Scientific Documentation, NSF - 61 - 76, November 1961.

26. Ramo-Wooldridge Division of Thompson-Ramo-Wooldridge Inc., Final Report on the Study of Automatic Abstracting, Contract No. AF 30(602)-2223, Canoga Park, California, September 1961.

27. Rockwell, H., "Information for Research and Development," presented at Symposium on the Relation between Management Information Systems and Information Retrieval, American Documentation Institute and Management Dynamics, New York, May 18-19, 1961.

28. Rockwell, H., R.L. Hayne, E. Garfield, "A Unique System for Rapid Access to Large Volumes of Pharmacologic Data," Federation Proceedings, Vol 16, No. 3, September 1957.

29. Spangler, M., General Bibliography on Information Storage and Retrieval, General Electric Company, 1962.

30. Thompson-Ramo-Wooldridge Inc., Work Correlation and Automatic Indexing - Phase II, a Final Report, Canoga Park, California, January 1962.

31. Waring, R.L., Technical Investigation of the Addition of a Hardcopy Output to Elements of a Mechanized Library System, Avco Corporation, Cincinnati, September 1961.

Aerospace Medical Division, 6570th Aerospace Medical Research Laboratories, Wright-Patterson AFB, Ohio. Rpt. No. AMRL-TDR-63-8, INFORMATION STORAGE AND RETRIEVAL: A SURVEY. Final report, Jan 63, iii + 15 pp., incl. 31 refs. Unclassified report

This report surveys the field of information retrieval, the methods of indexing and storing the vast number of scientific documents which have been produced in recent years. Information retrieval utilizes coordinate indexing—that is, listing documents under all the topics they contain and searching for them by two or more terms. There are two principle types of indexing:

(over)

UNCLASSIFIED

1. Indexes (Documentation)
2. Data Storage Systems (Computers)
3. Data Processing Systems (Computers)
4. Documentation
5. Bibliography (Documentation)
6. Abstracting
7. Subject Headings
8. Coding (Computers)
9. Punched Card Methods

UNCLASSIFIED

Aerospace Medical Division, 6570th Aerospace Medical Research Laboratories, Wright-Patterson AFB, Ohio. Rpt. No. AMRL-TDR-63-8, INFORMATION STORAGE AND RETRIEVAL: A SURVEY. Final report, Jan 63, iii + 15 pp., incl. 31 refs. Unclassified report

This report surveys the field of information retrieval, the methods of indexing and storing the vast number of scientific documents which have been produced in recent years. Information retrieval utilizes coordinate indexing—that is, listing documents under all the topics they contain and searching for them by two or more terms. There are two principle types of indexing:

(over)

UNCLASSIFIED

1. Indexes (Documentation)
2. Data Storage Systems (Computers)
3. Data Processing Systems (Computers)
4. Documentation
5. Bibliography (Documentation)
6. Abstracting
7. Subject Headings
8. Coding (Computers)
9. Punched Card Methods

UNCLASSIFIED

one using a predetermined list of terms into which all documents must be fitted and the other allowing free choice of the terms found in the documents themselves. Elaboration of these methods and the difficulty of developing a list of indexing terms are also discussed. An information retrieval system may consist of an index only, an index with an abstract, or an entire document with an index. The mechanical equipment used may range from punched cards through IBM cards to complex computers and microphotographic systems. The experiences of various organizations with different combinations of equipment and methods are discussed.

UNCLASSIFIED

I. Biomedical Laboratory
II. Barnard, G. W., Captain, USAF, MC
Abbott, C.
III. In ASTia collection
IV. Aval fr OTS: $0.75

UNCLASSIFIED

one using a predetermined list of terms into which all documents must be fitted and the other allowing free choice of the terms found in the documents themselves. Elaboration of these methods and the difficulty of developing a list of indexing terms are also discussed. An information retrieval system may consist of an index only, an index with an abstract, or an entire document with an index. The mechanical equipment used may range from punched cards through IBM cards to complex computers and microphotographic systems. The experiences of various organizations with different combinations of equipment and methods are discussed.

UNCLASSIFIED

I. Biomedical Laboratory
II. Barnard, G. W., Captain, USAF, MC
Abbott, C.
III. In ASTia collection
IV. Aval fr OTS: $0.75

UNCLASSIFIED